Predicting Emotion Perception Across Domains: A Study of Singing and Speaking

Biqiao Zhang, Emily Mower Provost, Robert Swedberg, Georg Essl
University of Michigan, Ann Arbor
2260 Hayward St.
Ann Arbor, Michigan 48109

Abstract

Emotion affects our understanding of the opinions and sentiments of others. Research has demonstrated that humans are able to recognize emotions in various domains, including speech and music, and that there are potential shared features that shape the emotion in both domains. In this paper, we investigate acoustic and visual features that are relevant to emotion perception in the domains of singing and speaking. We train regression models using two paradigms: (1) within-domain, in which models are trained and tested on the same domain and (2) cross-domain, in which models are trained on one domain and tested on the other domain. This strategy allows us to analyze the similarities and differences underlying the relationship between audio-visual feature expression and emotion perception and how this relationship is affected by domain of expression. We use kernel density estimation to model emotion as a probability distribution over the perception associated with multiple evaluators on the valence-activation space. This allows us to model the variation inherent in the reported perception. Results suggest that activation can be modeled more accurately across domains, compared to valence. Furthermore, visual features capture cross-domain emotion more accurately than acoustic features. The results provide additional evidence for a shared mechanism underlying spoken and sung emotion perception.

1 Introduction

Emotion expression and perception are vital components of social and musical communication (Cowie et al. 2001; Scherer 2003; Juslin and Sloboda 2001). Research in emotion perception has demonstrated that humans are able to recognize emotions in various domains, including speech and music. However, the relationship between audio-visual cues and emotion perception across domains is still an open question. In this paper, we aim to provide clarity by investigating the acoustic and visual features that are relevant to the perceived emotion of two types of vocal communications: singing and speaking. We conducted within-domain and cross-domain regression and feature correlation studies to analyze the commonalities and differences present in emotion expression across communication domains. Within-domain analyses build separate models for the singing and speaking domains, whereas cross-domain analyses focus on generalizing a model from one domain to the other. The ability of a model to generalize gives us insight into the link between emotion perception in the singing and speaking expression domains.

Significant progress has been made on both speech and music emotion recognition (Schuller et al. 2011; El Ayadi, Kamel, and Karray 2011; Kim et al. 2010). Some researchers have studied the similarity between music and speech emotion perception (Juslin and Laukka 2003; Ilie and Thompson 2006; 2011), yet few works (Scherer et al. 2013) have concentrated on comparing speaking and unaccompanied singing. Since music is generally only acoustically recorded, prior works primarily focus on acoustic features. Therefore, analysis on the role of visual features in cross-domain emotion perception is under-explored.

In this work, we collected a corpus of singing and speaking audio-visual recordings of three performers with experimental control over lexical content and melody to ensure the consistency of the data across communication types and target emotions. We assessed the emotion content using Amazon Mechanical Turk. We estimated the distribution of human evaluations using kernel density estimation to model emotion as a probability distribution on the valence-activation space (Yang and Chen 2011). We selected features relevant to within-domain and cross-domain emotion perception using mRMR (minimum Redundancy Maximum Relevance), with the correlation coefficient as the relevance measure for the former and the cross-domain correlation coefficient (Weninger et al. 2013) for the latter. We built regression models using the selected features for valence and activation, both individually and jointly as a two-dimensional emotion distribution, to assess the extent to which emotion perception can be predicted by the selected features.

The results demonstrated that activation can be modeled more accurately both within and across domains, compared to valence. Further, results suggest that visual features capture cross-domain emotion more accurately than acoustic features. The results provide additional evidence for a shared mechanism underlying spoken and sung emotion perception. The novelty of this paper includes: (1) the construction of the first dataset consisting of lexically and musically consistent

speaking and singing recordings with rich emotions using the same performers; (2) the introduction of visual features into cross-domain emotion perception analysis; (3) the extension of mRMR to cross-domain scenarios.

2 Related Works

2.1 Emotion Perception from Music

Many works have been done in music emotion recognition (MER) in recent years (Kim et al. 2010; Yang and Chen 2012). Most works on MER concentrated on acoustic features. Baume (2013) evaluated the acoustic features used in MER, and found that spectral features contribute the most to prediction accuracy. A very recent work demonstrated that the facial expressions of singers also influence the emotion perception of the audience (Quinto et al. 2014). Therefore, it is important to introduce visual features when analyzing emotion perception from singing expressions.

2.2 Emotion Perception from Speech

There have been studies looking at the emotion information encoded in audio (Schuller et al. 2011; Le and Mower Provost 2013), video (Metallinou et al. 2010; Kim and Mower Provost 2014) and audio-visual cues (Sebe et al. 2006; Mower, Mataric, and Narayanan 2011). Both acoustic and visual features are demonstrated useful for predicting speech emotion perception.

2.3 Cross-Domain Emotion Perception

Prior works have assessed the similarity between music and speech emotion perception. A meta-analysis of 104 studies of vocal expression and 41 studies of music performance found potential for a shared emotion perception mechanism between music and speech (Juslin and Laukka 2003). Ilie and Thompson (2006; 2011) found that the manipulation of certain acoustic features of music and speech, such as loudness and rate, results in similar emotion perception. Loud excerpts were judged as more pleasant, energetic, and tense. Fast music and speech were judged as having greater energy. Weninger et al. (2013) investigated the shared acoustic features in speech, music and sound domains and introduced the cross-domain correlation coefficient as a measure of relevance for cross-domain feature selection. Their research suggests that there are likely shared and divergent properties in emotion perception across domains of expression.

The comparison between emotion expressions of speaking and unaccompanied singing is relatively less explored. Research has indicated that high-level musical features, such as music structure and melodic complexity, influence emotion perception (Krumhansl and Agres 2008; Narmour 1992). This indicates that datasets that do not control for such variations may be influenced by these factors. Therefore, it is important to control these factors when studying the similarities and differences existing between non-musical and musical vocal emotion expression. Recently, a study comparing speaking and singing stimuli adopted spoken data recorded from French-speaking professional actors and sung data constructed from three professional opera singers (Scherer et al. 2013). The authors retained similarity in vocal content by using the same phrases for both types. They found a high degree of similarity in the value of acoustic parameters for energy, spectral flatness and jitter between singers' and actors' portrayals of emotion. However, factors that can also influence emotion expression, such as differences in performers and melodies, were not taking into account. Cross-domain emotion perception of speech and music has focused primarily on acoustic features (Juslin and Laukka 2003; Ilie and Thompson 2011; Weninger et al. 2013; Coutinho and Dibben 2013).

3 Dataset

3.1 Data Collection

We collected a corpus of speaking and singing performances. We recruited musical theater students that have completed coursework in the School of Music, Theater & Dance that included training in spoken and sung theatrical production. The finished dataset includes three performers (1 female, 2 male). The actors performed both domains in the same location under consistent visual and acoustic conditions. The vocal data were recorded via an Electro-Voice N/D 357 microphone and the video data were recorded using a high-definition Canon Vixia HF G10 camcorder.

Our dataset uses fixed lexical content. We identified seven semantically neutral sentences and embedded each sentence into four passages, each associated with a target emotion from the set of angry, happy, neutral and sad. This embedding allowed us to create an environment that would facilitate emotionally evocative performances. The consistency of the lexical content of the embedded target sentence allows for an analysis of emotion content while controlling for variation in lexical content. We composed seven stylistically neutral melodies in a singable range to match the seven passages for the singing performances. The target sentences were accompanied by the exact same melody. The remainder of the passage included minor differences across the four emotional variations to allow for differences in the lexical information. This resulted in 168 (2 domains of vocal expression $\times$ 3 performers $\times$ 7 sentences $\times$ 4 target emotions) excerpts in total. We segmented out the target sentence from the remainder of the passage for both speaking and singing performances. The average duration of the target sentences in the singing and speaking recordings are 3.04 ± 0.87 and 1.57 ± 0.37 seconds respectively.

3.2 Evaluation

We evaluated the target sentences using Amazon Mechanical Turk. The evaluation included the original audio-visual clips, only the audio information, and only the video information, which resulted in 504 utterances (168 $\times$ 3 types of stimuli). The evaluators assessed the emotion content across the dimensions of valence (positive/negative emotional states), activation (energy or stimulation level), and dominance (passive vs. dominant) (Russell 1980; Mehrabian 1980) using a 9-point Likert scale. The evaluators also assessed the primary emotion of the clips from the set of

angry, happy, neutral, sad and other. We only used evaluations of valence and activation in this work due to the high degree of correlation between the activation and dominance dimensions. We collected 10,531 evaluations in total, with 183 unique evaluators. Each utterance was evaluated by 20.9 ± 1.7 participants.

4 Methodology

4.1 Data Cleaning

The challenge of using human evaluation is to separate differences in opinions from noise. We used several methods to clean the evaluation data before further analysis.

We first removed the top rating and bottom rating of each clip for valence and activation. We then calculated the weighted kappa to identify evaluators whose evaluations are likely noise given the evaluations of other individuals. This method was demonstrated effective in (Mower Provost, Zhu, and Narayanan 2013). The weighted kappa between two evaluators A and B are given by

$$K = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{o,i,j}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{c,i,j}} \quad (1)$$

where $k = 9$ since we used a 9-point Likert scale, $w_{ij} = 2^{|i-j|}$ when i and j are different, $p_{o,i,j}$ is the observed probability of evaluator A choosing i and evaluator B choosing j, while $p_{c,i,j}$ is the chance probability. The score of each evaluator was calculated by taking the mean of the weighted kappa between this evaluator and all other evaluators who had assessed the same clips. We performed a z-test to identify outlier evaluators (α=0.05) (Grubbs 1969). We removed the two evaluators with z-score higher than the critical value. After that, we did z-normalization for each evaluator so that their evaluations have zero mean and standard deviation of one. For each utterance, we identified outlier evaluations using a z-test with a significance level of 0.05 over the valence and activation dimensions. Using these methods, the total number of evaluations was reduced to 8,540. Each utterance was evaluated by 16.9 ± 1.9 participants.

4.2 Dimensional Emotion Expression

We calculated the average valence and activation for each utterance. As emotion perception is by nature subjective, we worked with the distribution of evaluations, in addition to average evaluations. We applied density estimation to the Valence-Activation ("V-A") space to estimate the distribution of the individual evaluations for each utterance. This allows us to explicitly take the variation inherent in the reported perception into account. We used the kernel density estimation (KDE) to approximate the evaluator distribution. We used KDE instead of a method that first introduces binning grids and then counts the evaluations that fall into each grid to generate a density histogram. The histogram generated using the latter method highly depends on the position of the binning grids, which can lead to biasing. KDE was demonstrated effective in estimating the V-A distribution associated with popular music (Yang and Chen 2011).

We first found a continuous function for each utterance that approximates the distribution of the evaluations of that utterance (Figure 1(a)). In each utterance, KDE assigns energy to each of the human evaluations. In this way, the contribution of each evaluation is smoothed out from a single point into a region of space surrounding it. Aggregating the smoothed contributions gives an overall picture of the structure of the data and its density function, which we call the density ground truth ("DGT", Figure 1(b)). The function to calculate the density value of a position p is given by

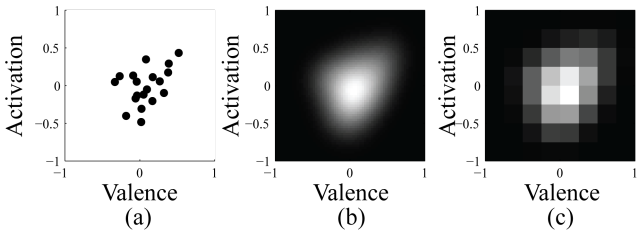


Figure 1: KDE. (a) distribution of evaluator judgment; (b) KDE approximation (DGT); (c) $G \times G$ grid-estimation.

$$y_i(p) = \frac{1}{E_i}\sum_{e=1}^{E_i} K(p - q_{ie}) \quad (2)$$

where E_i is the number of evaluators that annotated utterance i, q_{ie} is the evaluation of evaluator e on utterance i, and $K()$ is a bivariate Gaussian with zero mean and diagonal covariance (Botev et al. 2010).

Since the prediction of a continuous function is extremely challenging, we used a two-dimensional piecewise linear approximation of this continuous function. We created this approximation by drawing G equally spaced partitions across both valence and activation, resulting $G \times G$ individual grids (Yang and Chen 2011). The mean of the DGT values within each grid was used to represent the density of this grid (Figure 1(c)). We transform the 2D density estimation of each stimulus to a probability distribution by normalizing the value of $G \times G$ grids to sum to one.

4.3 Feature Extraction

We extracted audio-visual features to predict how individuals perceive the emotion in song and speech.

Acoustic Features We adopted the INTERSPEECH 2013 Computational Paralinguistics Evaluation (ComParE) feature set (Schuller et al. 2013). The set includes 4 energy, 41 spectral, 14 cepstral (MFCC) and 6 voicing-related low-level descriptors (LLDs). A variety of functions were applied to the LLDs and delta LLDs in order to summarize the evolution of the contours over time. We used OpenSMILE (Eyben, Wöllmer, and Schuller 2010) to extract the above 6,373 features.

Visual Features We extracted the visual features related to facial expression using CERT to get the estimation of the frame-by-frame intensity of 26 action units and 2 action unit groups (Littlewort et al. 2011). Action units are the fundamental actions of individual muscles or groups of muscles. Example facial action units that we used include inner/outer brow raise, eye widen, blink, lip corner pull, etc.

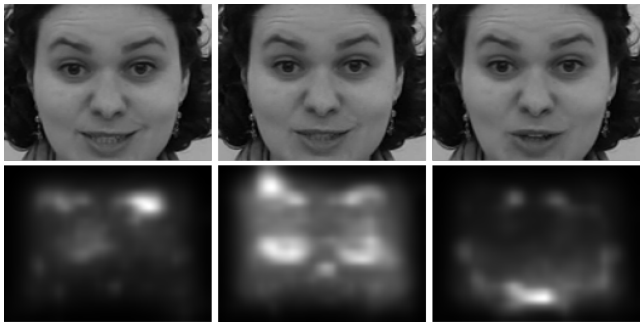


Figure 2: The face bounding box and corresponding saliency map for three consecutive frames. The saliency map changes with expression.

We are also interested in the distribution of visual saliency on the face because changes in visual saliency can cause visual attention to differ across emotions and faces (Scheller, Büchel, and Gamer 2012). Visual saliency ("saliency") is the perceptual distinctness that draws the viewer's attention to some parts of the face (Itti, Koch, and Niebur 1998). We calculated the saliency maps for all video stimuli. We first detected the face from the first frame of the video and tracked it in the following frames (Viola and Jones 2001; Tomasi and Kanade 1991; Kalal, Mikolajczyk, and Matas 2010). We then modified the bounding box to be the same size across frames and tilted it to be a rectangle that can be cropped from the original frame. Then, the whole face saliency of each frame was calculated using pixel intensity, flicker, and motion information of the current and previous frame (Harel, Koch, and Perona 2006). Researches have demonstrated that there are differences between emotion associated with the upper and lower face (Ross, Prodan, and Monnot 2007). Therefore, we divided the face into the upper and lower half and calculated the average saliency of the upper region, lower region, and the ration between the upper and lower saliency for each frame.

We applied statistics including mean, standard deviation, max, min, range, interquartile range, mean absolute deviation, skewness and kurtosis to action units, saliency, and delta of these contours to generate a description of these features for a whole video. This resulted in 558 features in total.

4.4 Feature Selection

We reduced the dimensionality of the feature set using mRMR (minimum Redundancy Maximum Relevance) for continuous variables. mRMR consists of two steps: (1) calculate the correlation between the target value and each feature to generate a pool of relevant features; (2) use the F-test correlation quotient to search for the best features in mRMR optimization conditions (Peng, Long, and Ding 2005).

We replaced the Pearson's correlation coefficient by the cross-domain correlation coefficient (CDCC) (Weninger et al. 2013) in stage (1) for cross-domain feature selection. This is the first method that extends mRMR to perform cross-domain tasks. The equation of CDCC is given by

$$CDCC_{f,i,j} = \frac{|r_f^{(i)} + r_f^{(j)}| - |r_f^{(i)} - r_f^{(j)}|}{2} \qquad (3)$$

where $r_f^{(i)}$ is the correlation of feature f with domain i.

The number of features selected was decided by optimizing over the training data. For audio-only, video-only and audio-visual stimuli, the numbers range from 50 to 300, 20 to 40 and 50 to 250 respectively.

4.5 Regression

We used $\nu - SVR$ with a radial basis function kernel implemented in Libsvm (Chang and Lin 2011). We performed two separate estimation tasks: (1) prediction of average valence/activation ratings and (2) prediction of the distribution of the evaluations in the V-A space. $\nu - SVR$ is a type of support vector regression that performs optimization using ν as the penalty parameter, where ν represents a lower bound on the fraction of samples that are support vectors.

We estimate the density distribution by training $G \times G$ regressors for each task. As $\nu - SVR$ is not constrained to output non-negative values, we transformed the output to a probability distribution by truncating negative values to zero and normalizing the estimation over all grids to sum to one.

4.6 Evaluation Method

We evaluated the models using leave-one-performer-out-cross-validation. For each round, we selected the recordings of two performers as the training set, and the remaining performer as the test set. The parameters are tuned on the training set using leave-one-utterance-out-cross-validation. We used the coefficient of determination (R^2) of the model when applied to the test set to measure the goodness of fit of the regression models. For density distribution, we calculated the R^2 between the DGT and the estimated emotion distribution for each utterance and then took the mean as the final result.

5 Results

5.1 Performance Study

We evaluated the performance of two regression tasks: (1) within-domain prediction, where models were tested on the same domain they were trained on; (2) cross-domain prediction, where models were trained on one domain but tested on the other. Domain-specific features and cross-domain features refer to features selected using Pearson's correlation and CDCC in the first stage of mRMR respectively.

We compared (1) the absolute prediction residual of individual utterance for valence/activation, and (2) the single utterance R^2 for density between models trained using domain-specific and cross-domain features for each prediction task using the paired t-test. The results is an indication of the significance of the difference in performance.

In within-domain predictions (Table 1(a)), models using domain-specific features achieve an R^2 of 0.94 for activation using both acoustic and visual features. All predictions of models trained with domain-specific features have an R^2 higher than 0.5 except for valence of the speaking domain. Similar to prior works (Yang et al. 2008), prediction of activation always outperforms valence, which indicates that activation is easier to estimate.

		Audio		Video		Aud-Vis	
	Domain	DSF	CDF	DSF	CDF	DSF	CDF
V	SI	0.57**	0.19	0.58	0.54	0.76**	0.52
	SP	0.75**	0.55	0.37*	0.21	0.65	0.66
A	SI	0.79	0.73	0.74	0.72	0.94**	0.85
	SP	0.95**	0.83	0.82**	0.58	0.94*	0.90
Den	SI	0.88**	0.80	0.71**	0.62	0.83**	0.74
	SP	0.86**	0.74	0.61**	0.54	0.83**	0.67

(a) Within-domain prediction R^2

			Audio		Video		Aud-Vis	
	Train	Test	DSF	CDF	DSF	CDF	DSF	CDF
V	SI	SP	0	0	0.29	0.44	0.14	0.38**
	SP	SI	0	0	0.35	0.24	0.18	0.29
A	SI	SP	0.08	0.36**	0.55	0.54	0.29	0.81**
	SP	SI	0.31	0.30	0.41	0.68**	0.54	0.53
Den	SI	SP	0.27	0.43**	0.46	0.52**	0.35	0.49**
	SP	SI	0.20	0.31**	0.51	0.60**	0.33	0.48**

(b) Cross-domain prediction R^2

Table 1: Performance of (a) within-domain and (b) cross-domain predictions. DSF: models using domain-specific features, CDF: models using cross-domain features, V: valence, A: activation, Den: density distribution, SI: singing, SP: speaking. ** and * mean that one model (DSF/CDF) is significantly better than the other in the same task under significance level of 0.01 and 0.05, respectively.

In cross-domain predictions (Table 1(b)), models trained with cross-domain features have significantly higher accuracy than those trained with domain-specific features in most cases. This suggests that by embedding CDCC into mRMR, we successfully increased the generalizability of cross-domain models. Estimation of activation is more accurate than valence. In addition, visual features work better than acoustic features or even combined audio-visual features. The difference between the performance of the cross-domain and within-domain prediction is the smallest in the video-only stimuli. This suggests that visual information is expressed similarly in both domains.

Models trained with domain-specific features outperform models trained with cross-domain features in within-domain prediction. Yet, these same models have lower accuracy in cross-domain prediction. This suggests that there is a trade-off between generalizability and domain-specific performance.

5.2 Analysis of Results

We grouped features by the type of their LLD and show their distribution in Figure 3 to study the audio-visual cues that contribute most to our ability to estimate perception. Spectral features and cepstral features are the most important acoustic features. Energy-related features are more relevant to activation than valence. The models trained to recognize video-only emotion perception are dominated by action unit features. Saliency features contain more information regarding valence than activation. Both acoustic and visual features play important roles in emotion perception from audio-visual stimuli, but emotion perception of the singing domain relies more heavily on visual information, especially for valence, compared to the spoken domain.

We conducted a feature correlation study to investigate the

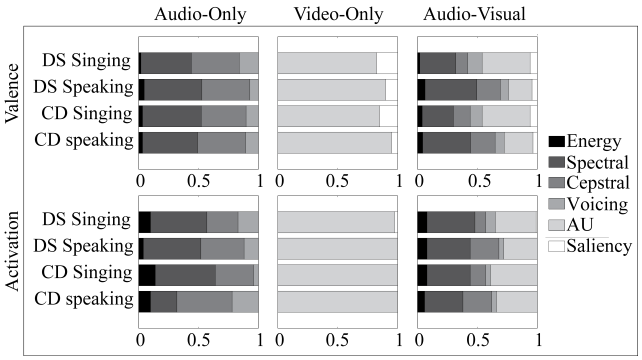


Figure 3: Features relevant to valence and activation of all types of stimuli. CD: cross-domain features; DC: domain-specific features. Features were grouped into 6 categories by Low Level Descriptor.

		Audio		Video		Aud-Vis	
	Domain	DSF	CDF	DSF	CDF	DSF	CDF
V	SI	0.31	0.23	0.40	0.39	0.37	0.30
	SP	0.28	0.24	0.33	0.34	0.28	0.22
A	SI	0.41	0.40	0.50	0.47	0.48	0.47
	SP	0.51	0.40	0.54	0.47	0.48	0.45

(a) Within-domain feature-emotion correlation

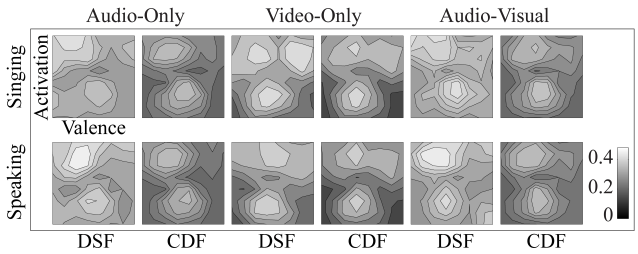
			Audio		Video		Aud-Vis	
	i	j	DSF	CDF	DSF	CDF	DSF	CDF
V	SI	SP	0.29	0.24	0.21	0.14	0.25	0.14
	SP	SI	0.28	0.22	0.17	0.09	0.20	0.15
A	SI	SP	0.22	0.19	0.21	0.12	0.24	0.14
	SP	SI	0.29	0.15	0.22	0.11	0.22	0.09

(b) Absolute difference in feature-emotion correlations across domains

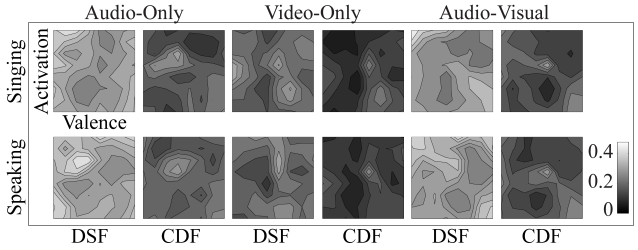
Table 2: (a) Within-domain average features-emotion correlations. (b) Cross-domain average $|r_f^{(song)} - r_f^{(speech)}|$. V: valence, A: activation, SI: singing, SP: speaking, DSF: domain-specific features, CDF: cross-domain features.

differences in performance between the domain-specific and cross-domain models on within-domain and cross-domain prediction tasks. Two types of correlations are calculated: (1) the Pearson's correlation coefficient between emotion perception and the selected features; (2) a sub-component of CDCC, $|r_f^{(song)} - r_f^{(speech)}|$, where $r_f^{(song)}$ is the correlation between feature f and the emotion perception in singing domain. The larger (1) is, the more relevant the features are; the larger (2) is, the less likely that the features are shared by the speaking and singing domains. Table 2 shows the results for average valence/activation prediction. Figure 4 visualizes the correlations between features and density distribution as 2D contour maps.

Table 2(a) shows that the correlation between domain-specific features and emotion perception is higher than that between cross-domain features and emotion perception. This corresponds to the performances of the feature sets on within-domain prediction. Similarly, Figure 4(a) shows that the correlation contours of domain-specific features



(a) Within-domain feature-emotion correlation contour



(b) Cross-domain correlation difference contour

Figure 4: In (a), lighter color means higher correlation. In (b), lighter color means higher difference between singing and speaking domain. DSF: Domain-specific features, CDF: cross-domain features.

are lighter, which means the feature-emotion correlation is higher compared to cross-domain features. The correlation for activation is always higher than for valence, which suggests that the acoustic and visual features we extracted are more relevant to the prediction of activation perception than valence perception. In figure 4(a), the regions with high and low activation are lighter in all contours, which indicates that the extreme values of activation perception can be more accurately estimated.

Table 2(b) shows that the absolute difference between feature-emotion correlations across domains in the valence dimension is larger for acoustic feautures, compared to visual features. This suggests that acoustic features have larger differences across domains. In contrast, this absolute difference in correlations across domains is small for cross-domain visual features. We can also observe this from Figure 4(b). The difference contours of cross-domain visual features are clearly darker, which indicate smaller differences across domains. We note that the difference in correlation is greater for domain-specific features across domains, compared to cross-domain features. This indicates that the cross-domain feature selection method we used is able to capture features shared by both domains.

6 Conclusion and Discussion

In this paper, we presented a novel cross-domain singing-speaking dataset. We estimated the emotion distribution on the valence-activation space using individual evaluations in addition to the conventional method of using average values. We selected features for domain-specific and cross-domain models using mRMR with two different relevant measures: Pearson's correlation coefficient and the CDCC. The latter one extended mRMR to cross-domain tasks. We built domain-specific and cross-domain models to predict valence, activation, and the density distribution of emotion across the V-A space. In within-domain prediction, we achieved the highest R^2 of 0.94 (audio-visual, activation) and only one lower than 0.5 (video, valence). In cross-domain prediction, models using cross-domain features outperformed models using domain-specific features, which indicated that we have successfully captured some similarity between speaking and singing domains.

Our results suggest that activation can be estimated more accurately across domains, compared to valence. Previous works have indicated that acoustic features capture activation information better than valence information for both music and speech (Mower Provost, Zhu, and Narayanan 2013; Yang et al. 2008). We found not only is activation more encoded in acoustic features, it is also more shared across expression domains, compared to valence.

Our results also demonstrated that visual features capture cross-domain emotion more accurately than acoustic features. The small differences in the performance between models using domain-specific features and models using cross-domain features in cross-domain prediction suggest that there may be a underlying shared perception model across the speaking and singing expression domains.

In this paper, we focused on utterance-level analysis. Previous works have shown that different phoneme groups are modulated differently by emotions (Busso, Lee, and Narayanan 2007; Kim and Mower Provost 2014). In our future work, we will examine how individual phonemes are shaped by emotion across domains.

One limitation of our work is size of the dataset. The small number of performers makes it hard to generalize. We plan to extend our dataset to include a larger number of performers and additional expression domains, such as instrumental music. In this way, we would be able to analyze the similarity and difference in emotion perception across speaking (non-musical, vocal), singing (musical, vocal) and instrumental music (musical, non-vocal) expression domains.

Acknowledgments

We would express our appreciation to Yuan Shangguan, who helped collect evaluations from Mechanical Turk and Isaac Levine, who composed the melodies of the sung stimuli.

References

Baume, C. 2013. Evaluation of acoustic features for music emotion recognition. In *Audio Engineering Society Convention 134*. Audio Engineering Society.

Botev, Z.; Grotowski, J.; Kroese, D.; et al. 2010. Kernel density estimation via diffusion. *The Annals of Statistics* 38(5):2916–2957.

Busso, C.; Lee, S.; and Narayanan, S. S. 2007. Using neutral speech models for emotional speech analysis. In *INTERSPEECH*, 2225–2228.

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3):27.

Coutinho, E., and Dibben, N. 2013. Psychoacoustic cues to emotion in speech prosody and music. *Cognition & Emotion* 27(4):658–684.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; and Taylor, J. G. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18(1):32–80.

El Ayadi, M.; Kamel, M. S.; and Karray, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44(3):572–587.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *International Conference on Multimedia*, 1459–1462. ACM.

Grubbs, F. E. 1969. Procedures for detecting outlying observations in samples. *Technometrics* 11(1):1–21.

Harel, J.; Koch, C.; and Perona, P. 2006. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, 545–552.

Ilie, G., and Thompson, W. F. 2006. A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception: An Interdisciplinary Journal* 23(4).

Ilie, G., and Thompson, W. F. 2011. Experiential and cognitive changes following seven minutes exposure to music and speech. *Music Perception: An Interdisciplinary Journal* 28(3).

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(11):1254–1259.

Juslin, P. N., and Laukka, P. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin* 129(5):770.

Juslin, P. N., and Sloboda, J. A. 2001. *Music and Emotion: Theory and Research.* Oxford University Press.

Kalal, Z.; Mikolajczyk, K.; and Matas, J. 2010. Forward-backward error: Automatic detection of tracking failures. In *Pattern Recognition*, 2756–2759. IEEE.

Kim, Y., and Mower Provost, E. 2014. Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition. In *ACM International Conference on Multimedia*.

Kim, Y. E.; Schmidt, E. M.; Migneco, R.; Morton, B. G.; Richardson, P.; Scott, J.; Speck, J. A.; and Turnbull, D. 2010. Music emotion recognition: A state of the art review. In *International Society for Music Information Retrieval*, 255–266. Citeseer.

Krumhansl, C. L., and Agres, K. R. 2008. Musical expectancy: The influence of musical structure on emotional response. *Behavioral and Brain Sciences* 31(05):584–585.

Le, D., and Mower Provost, E. 2013. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *Automatic Speech Recognition and Understanding*, 216–221. IEEE.

Littlewort, G.; Whitehill, J.; Wu, T.; Fasel, I.; Frank, M.; Movellan, J.; and Bartlett, M. 2011. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops*, 298–305.

Mehrabian, A. 1980. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Oelgeschlager, Gunn & Hain Cambridge, MA.

Metallinou, A.; Busso, C.; Lee, S.; and Narayanan, S. 2010. Visual emotion recognition using compact facial representations and viseme information. In *International Conference on Acoustics Speech and Signal Processing*, 2474–2477. IEEE.

Mower, E.; Mataric, M. J.; and Narayanan, S. 2011. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing* 19(5):1057–1070.

Mower Provost, E.; Zhu, I.; and Narayanan, S. 2013. Using emotional noise to uncloud audio-visual emotion perceptual evaluation. In *International Conference on Multimedia and Expo*, 1–6. IEEE.

Narmour, E. 1992. *The Analysis and Cognition of Melodic Complexity: The Implication-Realization Model*. University of Chicago Press.

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(8):1226–1238.

Quinto, L. R.; Thompson, W. F.; Kroos, C.; and Palmer, C. 2014. Singing emotionally: A study of pre-production, production, and post-production facial expressions. *Frontiers in Psychology* 5.

Ross, E. D.; Prodan, C. I.; and Monnot, M. 2007. Human facial expressions are organized functionally across the upper-lower facial axis. *The Neuroscientist* 13(5):433–446.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39(6):1161.

Scheller, E.; Büchel, C.; and Gamer, M. 2012. Diagnostic features of emotional expressions are processed preferentially. *PloS one* 7(7):e41792.

Scherer, K. R.; Sundberg, J.; Tamarit, L.; and Salomão, G. L. 2013. Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language*.

Scherer, K. R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40(1):227–256.

Schuller, B.; Batliner, A.; Steidl, S.; and Seppi, D. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 53(9):1062–1087.

Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E.; et al. 2013. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *INTERSPEECH*.

Sebe, N.; Cohen, I.; Gevers, T.; and Huang, T. S. 2006. Emotion recognition based on joint visual and audio cues. In *International Conference on Pattern Recognition*, volume 1, 1136–1139. IEEE.

Tomasi, C., and Kanade, T. 1991. *Detection and Tracking of Point Features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh.

Viola, P., and Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Conference on Computer Vision and Pattern Recognition*, volume 1, I–511. IEEE.

Weninger, F.; Eyben, F.; Schuller, B. W.; Mortillaro, M.; and Scherer, K. R. 2013. On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in Psychology* 4.

Yang, Y.-H., and Chen, H. H. 2011. Prediction of the distribution of perceived music emotions using discrete samples. *IEEE Transactions on Audio, Speech, and Language Processing* 19(7):2184–2196.

Yang, Y.-H., and Chen, H. H. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology* 3(3):40.

Yang, Y.-H.; Lin, Y.-C.; Su, Y.-F.; and Chen, H. H. 2008. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 16(2):448–457.